Time-varying Lasso

George Kapetanios, Filip Zikes

PII: S0165-1765(18)30166-6
DOI: https://doi.org/10.1016/j.econlet.2018.04.029
Reference: ECOLET 8028

To appear in: *Economics Letters*

Received date : 5 March 2018
Revised date : 23 April 2018
Accepted date : 24 April 2018

Please cite this article as: Kapetanios G., Zikes F., Time-varying Lasso. *Economics Letters* (2018), https://doi.org/10.1016/j.econlet.2018.04.029

*Highlights (for review)

Highlights

- We propose a novel lasso estimation for models with smoothly varying parameters
- The estimator is computationally simple and easy to implement in practice
- Methods for data-dependent choice of the regularization parameter are provided
- An application to forecasting inflation demonstrates the usefulness of the method

Time-varying Lasso*

George Kapetanios† Filip Zikes‡

April 23, 2018

Abstract

This paper introduces a Lasso-type estimator for large linear models with time-varying parameters. The estimator is easy to implement in practice and standard algorithms developed for Lasso with fixed parameters can be readily used. We derive theoretical properties of the estimator, allowing for deterministic or stochastic smoothly varying parameter processes and discuss ways in which tuning parameters can be data dependent. Monte Carlo simulation and an application to forecasting inflation with macroeconomic variables illustrates the usefulness of our method.

JEL Classification: C32, C52, E37
Keywords: Large Datasets, Structural Change, Penalised Regressions, Lasso.

1 Introduction

The issue of specifying correctly a model has been a major preoccupation in econometrics and statistics. The recent explosion in the availability of large datasets has made this specification task considerably harder. In particular, the presence of datasets where the number of potential regressors for a given regression model can be of the same or larger order of magnitude compared to the number of observations, has spurred considerable advances in statistical and econometric methodology. Model selection and estimation in this high-dimensional setting has largely settled around a set of methods collectively known as penalised regression. Penalised regression is an extension of multiple regression where the vector of regression coefficients, β^0 of a regression of y_t on $x_t = (x_{1,t}, ..., x_{N,t})'$ is estimated by $\hat{\beta}$ where $\hat{\beta} = \arg\min_{\beta} \left[\sum_{t=1}^{T}(y_t - x_t'\beta)^2 + Q(\beta, \lambda)\right]$, in which $Q(\beta, \lambda)$ is a penalty function that penalises the complexity of β, while λ is a vector of tuning parameters to be set by the researcher. A selective list of seminal contributions to this

*The views expressed in this paper are those of the authors and not necessarily those of the Federal Reserve Board or any other person associated with the Federal Reserve System.

†King's College, London.

‡Federal Reserve Board, Division of Financial Stability, 20th Street and Constitution Avenue N.W., Washington, D.C., 20551, United States. Phone: +1-202-475-6617. Email: filip.zikes@frb.gov.

literature includes Tibshirani (1996), Zhou and Hastie (2005), Lv and Fan (2009), Efron et al. (2004), Bickel et al. (2009), Fan and Li (2001) and Antoniadis and Fan (2001).

Throughout this extensive work programme on penalised regressions, which has been mainly initiated and developed in the statistical literature, little emphasis has been placed on exploring the implications of the possibility that observations are serially dependent and, even less, that this dependence changes across observations. In this paper we explore penalised regression within a time series context, focusing for simplicity on Lasso, in the presence of structural change. There is little theoretical work on the properties of Lasso under structural change or on possible modifications to Lasso to account for structural change. A simple solution that has been used in practice is the use of rolling windows. Raviv and van Dijk (2013), Li and Chen (2014), and Demirer et al. (2018) are recent examples, but they are purely empirical implementations and do not discuss any theoretical issues. Lee et al. (2014) discuss the theoretical properties of Lasso in the presence of a single change-point which, while theoretically interesting, is not empirically relevant in the presence of sustained structural change. As a result we consider a different theoretical paradigm of structural change.

We follow the work of Giraitis et al. (2014) where parameters are assumed to shift smoothly and possibly randomly. Under the assumption that the parameter processes are persistent and bounded, they show that they can be estimated consistently using kernel estimation. Following on from this paradigm we analyse a kernel Lasso estimator. We derive consistency and rate results both for finite and large number of covariates. Further, we undertake extensive Monte Carlo and empirical exercises which confirm the usefulness and applicability of our proposed methods. The rest of the paper is structured as follows: Section 2 presents our theoretical results and Sections 3 and 4 report Monte Carlo and empirical evidence in support of our proposed estimators. Proofs are relegated to an Appendix.

2 Theory

2.1 Theoretical results for finite N

We first report results for the case where the number of covariates is assumed finite. Our regression model is given by $y_t = x_t'\beta_t^0 + \epsilon_t$,and the Lasso estimator is given by $\hat{\beta}_t = \arg\min_{\beta\in\Omega}\left\{\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(y_j - x_j'\beta\right)^2 + \frac{\lambda_H}{H}\left\|\beta\right\|_1\right\}$, where w_{tj} are weights to be specified below and H is a bandwidth parameter. We make the following assumptions.

Assumption 1 *$x_t = (x_{1,t}, ..., x_{N,t})'$ is a vector of heterogeneous $\alpha-$mixing processes with mixing coefficients, α_{ik}, satisfying $\sup_i \alpha_{ik} \leq C\xi^k$ for some $C < \infty$, $0 < \xi < 1$.*

$Var(x_t) = \Sigma$, *full rank.* ϵ_t *is a heterogeneous* $\alpha-$*mixing process,* $E\left(\epsilon_t^2\right) = \sigma^2$, $E(x_t\epsilon_t) = 0$ *and* $E(x_{i,t}\epsilon_t x_{j,s}\epsilon_s) = 0$ *for all* i, j, t, s, $t \neq s$.

Assumption 2 $\beta_t^0 = (\beta_{1,t}^0, ..., \beta_{N,t}^0)'$ *is a bounded random/deterministic process independent of* ϵ_t *satisfying*

$$\sup_{|j|\leq h} \left\| \left(\beta_{t-j}^0 - \beta_t^0\right) \right\|^2 = O_p\left(\frac{h}{t}\right), \tag{1}$$

as $t \to \infty$, $h \to \infty$, $h = o(t)$.

Assumption 3 *Let* $w_{tj} := \frac{H\tilde{w}_{tj}}{\sum_{j=1}^T \tilde{w}_{tj}}$, $\tilde{w}_{tj} = K(\frac{t-j}{H})$, $H = o(T)$, $H \to \infty$, $K(x) \geq 0$, $x \in \mathbb{R}$ *is a continuous bounded function (kernel) with a bounded first derivative such that* $\int K(x)dx = 1$. *Further,* $K(x) = O(e^{-cx^2})$, $\exists c > 0$, $|(d/dx)K(x)| = O(|x|^{-2})$, $\quad x \to \infty$.

Remark 1 *(1) of Assumption 2 is not standard. A more standard assumption from the time varying statistical literature would replace the RHS of (1) with* $O_p\left(\frac{h}{T}\right)$. *Nevertheless, there is a number of reasons why that is suboptimal. An obvious first point is that if* $\sup_{|j|\leq h} \left\| \left(\beta_{t-j}^0 - \beta_t^0\right) \right\|^2 = O_p\left(\frac{h}{T}\right)$, *then (1) is also satisfied and, therefore, the standard assumption is stricter. Secondly, the direction of time makes the assumption rather suspect. An extra observation under the standard assumption implies increased smoothness over the whole sample rather than just the end of the sample. Examples of* β_t^0 *are discussed in Giraitis et al. (2018).*

Then, the following Theorem proves consistency of the estimator.

Theorem 1 *Let Assumptions 1-3 hold,* $y_t = x_t'\beta_t^0 + \epsilon_t$ *and* $\hat{\beta}_t = \arg\min_{\beta\in\Omega} \left\{ \begin{array}{c} \frac{1}{H}\sum_{j=1}^T w_{tj}\left(y_j - x_j'\beta\right)^2 + \\ \frac{\lambda_H}{H}\left\|\beta\right\|_1 \end{array} \right\}$ *where* Ω *is a compact subset of* R^N. *Then for all* $t = [\tau T]$, $0 < \tau < 1$, *if* $H \to \infty$, $(H\log^{1/2} H)/T \to 0$ *as* $T \to \infty$ *and* $\lambda_H = o(H)$, $\hat{\beta}_t \xrightarrow{p} \beta_t^0$.

2.2 Large n

Next, we move on to a setting where the number of covariates can tend to infinity, potentially at a faster rate than the number of observations. The following slightly stronger set of assumptions is made.

Assumption 4 $x_t = (x_{1,t}, ..., x_{N,t})'$ *is a vector of heterogeneous* $\alpha-$*mixing processes with mixing coefficients given by* $\sup_i \alpha_{ik} \leq C\xi^k$ *for some* $C < \infty$, $0 < \xi < 1$. ϵ_t *is a heterogeneous* $\alpha-$*mixing processes with mixing coefficients given by* $\sup_i \alpha_{ik} \leq C\xi^k$ *for all* $C < \infty$, $0 < \xi < 1$. $E\left(x_t\epsilon_t\right) = 0$. x_t *and* ϵ_t *satisfy*

$$\sup_i \Pr\left[|x_{i,t}| > a\right] \leq C_1 e^{-C_2 a^q},\ q > 1 \text{ and } \Pr\left[|\epsilon_t| > a\right] \leq C_1 e^{-C_2 a^q},\ q > 1, \tag{2}$$

for some $C_1, C_2 > 0$.

Assumption 5 *Let* $x_t = (x_{1,t}, ..., x_{N,t})'$ *and* $\beta_t^0 = (\beta_{1,t}^0, ..., \beta_{N,t}^0)'$. *Then,* $\sup_{|j|\leq h} \left\| x_{t-j}' \left(\beta_{t-j}^0 - \beta_t^0 \right) \right\| = O_p\left(\frac{h}{t}\right)$, *and* $E\left|x_t'\beta_t^0\right|^2 < \infty$.

Then the following Theorem provides an upper bound for the norm of the estimation error in terms of the norm of the true regression coefficient. This provides both consistency and a rate of convergence under assumptions for the latter norm.

Theorem 2 *Let Assumptions 4-5 and 3,* $y_t = x_t'\beta_t^0 + \epsilon_t$ *and* $\hat{\beta}_t = \arg\min_\beta \left\{ \begin{array}{c} \frac{1}{H}\sum_{j=1}^T w_{tj}\left(y_j - x_j'\beta\right)^2 + \\ \lambda \left\|\beta\right\|_1 \end{array} \right\}$ *where* $\lambda = \lambda_T = (\log N)^2 H^{-\frac{1}{2}} + \frac{H}{T}$. *Then, with probability approaching 1, as* $N, T \to \infty$, $\frac{1}{H}\sum_{j=1}^T w_{tj}\left(x_j'\left(\beta_t^0 - \hat{\beta}_t\right)\right)^2 \leq 3\lambda \left\|\beta_t^0\right\|_1$.

Of course this is just a result for a simple Lasso implementation but it clearly illustrates the way to prove results in our time varying context and thus enables results to be developed for more sophisticated penalised regression estimators.

2.3 Data-driven choice of λ^*

The theory presented above provides guidance for the choice of the bandwidth, H, but it does not suggest how to choose λ in practice. Here we outline a simple cross-validation approach that can be used to this end, and is taylored to our time series setting. We distinguish between the in-sample estimation, where we typically use a two-sided kernel, and out-of-sample prediction, which requires a one-sided kernel. For In-sample parameter estimation we proceed as follows: We set up a grid for the shrinkage parameter $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_K\}$ and estimate for each λ in Λ the parameter vector β_t by our kernel estimator *leaving out the t-th observation*. We denote this leave-t-out estimator by $\hat{\beta}_{-t}(\lambda)$. Then, for computational parsimony, we only estimate the βs for all t on some grid $\tau = \{t_1, t_2, ..., t_M\}$, $M < T$. We suggest taking $t_1 = H$ and $t_M = T - H$. Since $H = T^{1/2}$, this grid covers an increasing fraction of $[1, T]$ as $T \to \infty$. In the simulations reported below, the increments $t_i - t_{i-1}$ vary between 2 and 8 depending on the sample size T. Having calculated $\hat{\beta}_{-t}(\lambda)$ for all $t \in \tau$ and $\lambda \in \Lambda$, we then choose the optimal lambda according to $\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \sum_{t\in\tau}(y_t - \hat{\beta}_{-t}(\lambda)'x_t)^2$. The same $\hat{\lambda}$ is then used to estimate β_t for all $t \in \tau$.

For out-of-sample prediction, the two-sided kernel is infeasible. We can use a one-sided kernel to estimate β_T and construct the forecast $\tilde{\beta}_T' x_{T+1}$, where $\tilde{\beta}_T$ is a feasible estimator of β_T. We again take a grid Λ and a grid τ, where we now set $\tau = \{T - M, T - M + 1, ..., T - 1\}$ for some $M < T$, and estimate β_t for each $t \in \tau$ and each

$\lambda \in \Lambda$. We denote this estimator by $\tilde{\beta}_t(\lambda)$. We then choose the optimal λ according to $\tilde{\lambda} = \arg\min_{\lambda \in \Lambda} \sum_{t \in \tau} (y_{t+1} - \tilde{\beta}_t(\lambda)' x_{t+1})^2$. Finally, we obtain the feasible estimator $\tilde{\beta}_T = \tilde{\beta}_T(\tilde{\lambda})$ and a feasible forecast $\tilde{\beta}_T' x_{T+1}$.

3 Monte Carlo simulation

In this section, we run a Monte Carlo experiment to investigate the properties of our kernel estimator in small samples. We peform two sets of experiments. First, we focus on the in-sample performance of the time-varying lasso, employing the standard Gaussian kernel, and calculate the MSE of $\hat{\beta}_t$. Second, we employ a one-sided version of the Gaussian kernel and only use past data to estimate the current parameter value, which we then use to construct one-step-ahead forecasts of y_{T+1} as discussed in Section 2.3. In both cases, we compare the performance of the time-varying lasso with the standard full-sample lasso.

The Monte Carlo design is as follows. We consider a linear regression model with N covariates:

$$y_t = \beta_t' x_t + \sqrt{\kappa} u_t, \quad u_t \overset{iid}{\sim} \mathrm{N}(0,1), \tag{3}$$

where the parameter process follows a scaled random walk,

$$\beta_{jt} = \frac{1}{\sqrt{t}} \tilde{\beta}_{jt}, \quad \tilde{\beta}_{jt} = \tilde{\beta}_{j,t-1} + \eta_{jt}, \quad \eta_t \overset{iid}{\sim} \mathrm{N}(0, I), \quad j = 1, ..., 4, \tag{4}$$

and $\beta_{jt} \equiv 0$, $j = 5, ..., N$, and the covariates obey

$$x_{1t} = \epsilon_{1t}, \quad x_{jt} = (\epsilon_{j-1,t} + \epsilon_{jt})/\sqrt{2}, \quad j = 2, 3, 4, \tag{5}$$

$$x_{5t} = \epsilon_{5t}, \quad x_{jt} = (\epsilon_{j-1,t} + \epsilon_{jt})/\sqrt{2}, \quad j = 6, 7, ..., N, \tag{6}$$

where the ϵ's are mutually independent Gaussian AR(1) processes with unit variance and autoregressive parameter equal to either 0 (*iid*) or 0.75 (persistent regressors). Thus the covariates are mutually dependent, potentially serially correlated, zero-mean, unit-variance processes. Only the first four covariates have a non-zero slope coefficients in (3) (almost surely). The parameter processes for the first four regressors follow independent random walks, scaled by t to satisfy condition (x). We set κ equal to 2, 4, or 10, implying a signal-to-noise ratio of 3.71, 1.85, 0.74, respectively. We consider model dimensions of $N = 50$ and sample sizes of $T = 200, 400, 800, 1600$ and 3200 observations.

A few remarks on the implementation of the time-varying lasso are in order. First, we set the kernel bandwidth according to $H = T^{1/2}$ as suggested by theory. Second, we employ the cross-validation procedures outlined in the previous section to select

the optimal shrinkage parameter λ. The grid for λ is set following Friedman et al. (2010). Finally, to reduce the computational burden, we restrict the number of non-zero parameter estimates, such that for any N the maximum number of non-zero βs at any given point in time t does not exceed 10.

The simulation results for the first set of experiments are reported in rows labeled "In-sample" in Table 1. To save space, we only report result for *iid* covariates, noting that the results for serially correlated covariates are qualitatively similar and available upon request. In general, we find that the kernel lasso performs better than the standard lasso for higher values of the signal-to-noise ratio and for larger samples. In these cases, there are significant improvements in the MSE relative to the fixed-parameter lasso. The results for the second set of experiments are reported in the rows labeled "Out-of-sample" in Table 1. Similar to the previous simulations, we find that significant gains in forecasting accuracy are obtained by using the kernel lasso in situations where the sample size is large and the population R^2 is relatively high.

4 Empirical illustration

To illustrate the usefulness of our methodology, we now present an empirical application to forecasting inflation. We focus on simple linear models where inflation depends on its own lag and on 25 lagged macroeconomic indicators listed in the Appendix. We obtain the monthly time-series of these macroeconomic variables from the FRED-MD database of U.S. macroeconomic indicators and apply the transformations suggested by McCracken and Ng (2015). Our sample period runs from January 1959 to June 2015 and we reserve the last 250 months for out-of-sample forecast evaluation. Results based on 200 and 300 out-of-sample observations are qualitatively similar.

To assess out-of-sample performance, we calculate the mean square error (MSE) and the mean absolute devation (MAD) for one-step-ahead forecasts generated by the time-varying lasso, which we implement using the approach discussed in subsection 2.3, and the standard lasso, where we adopt the recursive scheme (i.e. we always use all available data to generate the feasible one-step-ahead forecasts) and 10-fold cross validation. For comparison, we also calculate the MSE and MAD for the time-varying AR(1) model of Giraitis et al. (2014) and the standard AR(1). We use the Diebold-Mariano test for pairwise comparison of predictive ability of the different models.

The results are reported in Table 2. We find that our the time-varying lasso delivers the most accurate forecasts both in terms of the MSE and MAD. In the latter case, the improvements over standard lasso and the AR(1) are statistically significant at the 5% level using the Diebold-Mariano test for equal predictive accuracy. Although the time-varying AR(1) performs worse than the time-varying lasso, the two model forecasts

are statistically indistinguishable.

References

Antoniadis, A. and J. Fan, 2001, Regularization of Wavelets Approximations, *Journal of the American Statistical Association*, 96, 939–967.

Bickel, J. P., Ritov, Y. and A. Tsybakov, 2009, Simultaneous Analysis of Lasso and Dantzig Selector, *Annals of Statistics*, 37, 1705–1732.

Buhlmann, P. and S. van de Geer, 2011, *Statistics for High-dimensional Data: Methods, Theory and Applications*, New York: Springer.

Demirer, M., Diebold, F.X., Liu, L. and K. Yilmaz, 2018, Estimating Global Bank Network Connectedness, *Journal of Applied Econometrics*, 33(1), 1–15.

Efron, B., Hastie, T., Johnstone, I. and R. Tibshirani, 2004, Least Angle Regression, *Annals of Statistics*, 32, 407–499.

Fan, J. and R. Li, 2001, Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties, *Journal of the American Statistical Association*, 96, 1348–1360.

Friedman, J., Hastie, T. and R. Tibshirani, 2010, Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33(1), 1–22.

Giraitis, L., Kapetanios, G., Zikes, F., and A. Wetherilt, 2016, Estimating the dynamics and persistence of financial networks, with an application to the sterling money market, *Journal of Applied Econometrics*, 31(1), 58–84.

Giraitis, L., Kapetanios, G., and T. Yates, 2014, Inference on Stochastic Time-varying Coefficient Models, *Journal of Econometrics*, 179, 46–65.

Giraitis, L., Kapetanios, G., and T. Yates, forthcoming, Inference on multivariate heteroscedastic time varying random coefficient models, *Journal of Time Ser. Anal.*

Knight, K. and W. Fu, 2000, Asymptotics for LASSO-type Estimators, *Annals of Statistics*, 28(5), 1356-1378.

Lee, S., Seo, M.H. and Y. Shin, 2016, The LASSO for High-dimensional Regression with a Possible Change Point, *Journal of the Royal Stat. Soc. Ser. B*, 78(1), 193–210.

Li, J. and W. Chen, 2014, Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models, *International Journal of Forecasting*, 30, 996-1015.

Lv, J. and Y. Fan, 2009, A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares, *Annals of Statistics*, 37, 3498–3528.

McCracken, M. and S. Ng, 2015, FRED-MD: A Monthly Database for Macroeconomic Research, Working Paper 2015-012B, Federal Reserve Bank of St. Louis.

Raviv, E. and D. van Dijk, 2013, Forecasting with Many Predictors: allowing for nonlinearity, mimeo, Erasmus University Rotterdam.

Tibshirani, R., 1996, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1), 267–288.

Zhou, H. and T. Hastie, 2005, Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society B*, 67, 301–320.

A Variables used in the forecasting exercise

CPI:All Items; Real Pers. Inc.; IP Index; Capacity Util.: Manufact.; Civ. Unemp. Rate; Avg Hourly Earnings: Goods-Prod.; Avg Hourly Earnings: Constr.; Avg Hourly Earnings: Manufact.; Housing Starts; ISM: PMI Composite Index; ISM: New Orders Index; ISM: Inventories Index; M1 Money Stock; M2 Money Stock; St. Louis Adj. Mon. Base; Commercial and Industrial Loans; Effective Fed. Funds Rate; 1-Year Treasury Rate; 10-Year Treasury Rate; Moody's Baa Corp. Bond Minus Fed. Funds Rate; PPI: Finished Goods; PPI: Intermed. Materials; PPI: Crude Materials; Crude Oil, spliced WTI and Cushing; Pers. Cons. Expend.: Chain Index; S&P 500: Composite;

B Proofs of Theorem 1 and 2

Proof of Theorem 1. Following Knight and Fu (2000) it suffices to show that $\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(y_j - x_j'\beta\right)^2$ converges in probability to $(\beta_t^0 - \beta)'Q(\beta_t^0 - \beta) + \sigma^2$ pointwise in Ω, where Q is some regular matrix. Write

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(y_j - x_j'\beta\right)^2 = \frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'(\beta_j^0 - \beta)\right)^2 + \frac{2}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j x_j'(\beta_j^0 - \beta) + \frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j^2 = \sum_{j=1}^{3} A_{j,T,H} \tag{7}$$

Then

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'(\beta_j^0 - \beta)\right)^2 = \frac{3}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'(\beta_j^0 - \beta_t^0)\right)^2 + \tag{8}$$

$$\left(\frac{2}{H}\sum_{j=1}^{T} w_{tj}x_j'(\beta_j^0 - \beta_t^0)x_j'\right)(\beta_t^0 - \beta) + (\beta_t^0 - \beta)'\left(\frac{1}{H}\sum_{j=1}^{T} w_{tj}x_j x_j'\right)(\beta_t^0 - \beta)$$

Now

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'(\beta_j^0 - \beta_t^0)\right)^2 = \frac{1}{H}\sum_{j:|j-t|<h} w_{tj}\left(x_j'(\beta_j^0 - \beta_t^0)\right)^2 + \frac{1}{H}\sum_{j:|j-h|\geq h} w_{tj}\left(x_j'(\beta_j^0 - \beta_t^0)\right)^2 \tag{9}$$

where we set $h = bH\log^{1/2} H$ for some $b > 0$ such that $\frac{1}{H}\sum_{j:|j-h|\geq h} w_{tj} = o(H^{-1})$ (Giraitis et al., 2016). Then since β_t^0 is bounded and the second moments of x_t exist, by Markov inequality the second term on the right-hand side of (9) is $o_p(H^{-1})$. Turning to

the first term, we have

$$\frac{1}{H}\sum_{j:|j-t|<h} w_{tj}\left(x_j'(\beta_j^0-\beta_t^0)\right)^2 = \frac{1}{H}\sum_{j:|j-t|<h} w_{tj}\left(\sum_{i=1}^N x_{i,j}(\beta_{i,j}^0-\beta_{i,t}^0)\right)^2 \leq$$

$$\sup_{j:|j-t|<h}\|\beta_j^0-\beta_t^0\|^2\left(\frac{1}{H}\sum_{j:|j-t|<h} w_{tj}\left(\sum_{i=1}^N x_{i,j}^2\right)\right) = O_p\left(\frac{H\log^{1/2}H}{T}\right)$$

since by Assumptions 1 and 3, $\frac{1}{H}\sum_{j:|j-t|<h} w_{tj}\left(\sum_{i=1}^p x_{i,j}^2\right) = O_p(1)$. Next we handle the second term on the right-hand side of (8). By the same argument as above, it suffices to look at the sum over j such that $|j-t|<h$. We have for any i, $i=1,...,N$,

$$\left|\frac{1}{H}\sum_{j:|j-t|<h} w_{tj}x_{i,j}x_j'(\beta_j^0-\beta_t^0)\right| \leq \max_{j:|j-t|<h}\|\beta_j^0-\beta_t^0\|\left(\frac{1}{H}\sum_{j:|j-t|<h} w_{tj}|x_{i,j}|\left(\sum_{k=1}^N x_{k,j}^2\right)^{1/2}\right) =$$

$$O_p\left(\left(\frac{H\log^{1/2}H}{T}\right)^{1/2}\right)$$

since $\|\beta_j^0-\beta_t^0\| \leq \left(\sup_{j:|j-t|<h}\|\beta_j^0-\beta_t^0\|^2\right)^{1/2}$ for all j, $|j-t|<h$, implies that $\max_{j:|j-t|<h}\|\beta_j^0-\beta_t^0\| = O_p\left((h/t)^{1/2}\right)$. Finally, by Assumptions 1 and 3, $\frac{1}{H}\sum_{j=1}^T w_{tj}x_jx_j'$ converges in probability to Σ (at rate $H^{1/2}$). In summary, if $\frac{H\log^{1/2}H}{T}\to 0$ as $H,T\to\infty$, $A_{1,T,H}$ converges in probability to $(\beta_t^0-\beta)'\Sigma(\beta_t^0-\beta)$.

Turning to $A_{2,T,H}$, note that by Assumption 1 the summands have zero expectation and are serially uncorrelated. This implies $E(A_{2,T,H})=0$ and $E(A_{2,T,H}^2) = \sum_{j=1}^T w_{tj}^2 E\left(\epsilon_j^2(x_j'(\beta_j^0-\beta))^2\right) \leq C\sum_{j=1}^T w_{tj}^2 E(\epsilon_k^2)E\left(\sum_{i=1}^N x_{i,j}^2\right) = O(H^{-1})$, since β_t^0 is bounded. Hence, $A_{2,T,H}=O_p(H^{-1/2})$. Finally, by the LLN, $A_{3,T,H}=\sum_{j=1}^T w_{tj}\epsilon_j^2 \xrightarrow{p} \sigma^2$ proving the result. ■

Proof of Theorem 2. We proceed in a number of steps. The first is the equivalent of the basic inequality (Lemma 6.1 of Buhlmann and van de Geer (2011), (BgV)). The first thing to note is that

$$\frac{1}{H}\sum_{j=1}^T w_{tj}\left(y_j-x_j'\hat\beta_t\right)^2+\lambda\left\|\hat\beta_t\right\|_1 \leq \frac{1}{H}\sum_{j=1}^T w_{tj}\left(y_j-x_j'\beta_t^0\right)^2+\lambda\left\|\beta_t^0\right\|_1 \tag{10}$$

Then

$$\frac{1}{H}\sum_{j=1}^T w_{tj}\left(y_j-x_j'\hat\beta_t\right)^2 = \frac{1}{H}\sum_{j=1}^T w_{tj}\left(x_j'\left(\beta_j^0-\hat\beta_t\right)+\epsilon_j\right)^2 =$$

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)^2+\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_t^0-\hat{\beta}_t\right)\right)^2+\frac{2}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)\left(x_j'\left(\beta_t^0-\hat{\beta}_t\right)\right)+$$

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j^2+\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j x_j'\left(\beta_j^0-\beta_t^0\right)+\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j x_j'\left(\beta_t^0-\hat{\beta}_t\right)=\sum_{j=1}^{6} A_i$$

Also, $\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(y_j-x_j'\beta_t^0\right)^2=\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)^2+\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j^2-\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j x_j'\left(\beta_j^0-\beta_t^0\right)$ $\sum_{j=1}^{3} B_i$. It is clear that $A_1=B_1$, $A_4=B_3$ and $A_5=B_3$. So (10) implies

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_t^0-\hat{\beta}_t\right)\right)^2+\frac{2}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)\left(x_j'\left(\beta_t^0-\hat{\beta}_t\right)\right)+ \quad (11)$$

$$\lambda\left\|\hat{\beta}_t\right\|_1\leq\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j x_j'\left(\hat{\beta}_t-\beta_t^0\right)+\lambda\left\|\beta_t^0\right\|_1 \quad (12)$$

Then, we proceed as follows:

$$\frac{1}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_t^0-\hat{\beta}_t\right)\right)^2+\lambda\left\|\hat{\beta}_t\right\|_1$$
$$\leq\left(\frac{1}{H}\sum_{j=1}^{T} w_{tj}\epsilon_j x_j'\right)\left(\hat{\beta}_t-\beta_t^0\right)-\left(\frac{2}{H}\sum_{j=1}^{T} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)x_j'\right)\left(\hat{\beta}_t-\beta_t^0\right)+\lambda\left\|\beta_t^0\right\|_1$$

But, given assumption 3, $\frac{1}{H}\sum_{|j-t|>H} w_{tj}\epsilon_j x_{i,j}=o_p(1)$ and $\frac{2}{H}\sum_{|j-t|>H} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)x_j'=o_p(1)$. Then we have

$$\left(\frac{1}{H}\sum_{|j-t|\leq H} w_{tj}\epsilon_j x_j'\right)\left(\hat{\beta}_t-\beta_t^0\right)-\left(\frac{2}{H}\sum_{|j-t|\leq H} w_{tj}\left(x_j'\left(\beta_j^0-\beta_t^0\right)\right)x_j'\right)\left(\hat{\beta}_t-\beta_t^0\right)+$$

$$\lambda\left\|\beta_t^0\right\|_1\leq\max_{i\leq N}\left(\frac{1}{H}\left|\sum_{|j-t|\leq H} w_{tj}\epsilon_j x_{i,j}\right|\right)\left\|\hat{\beta}_t-\beta_t^0\right\|_1+$$

$$\max_{|j-t|\leq H}\left\|x_j'\left(\beta_j^0-\beta_t^0\right)\right\|\left[\begin{array}{c}\max_{i\leq N}\left(\frac{2}{H}\left|\sum_{|j-t|\leq H} w_{tj}\left(|x_{i,j}|-E\left(|x_{i,j}|\right)\right)\right|\right)+\\ \max_{i\leq N}\left(\frac{2}{H}\sum_{|j-t|\leq H} w_{tj}E\left(|x_{i,j}|\right)\right)\end{array}\right]\times\left\|\hat{\beta}_t-\beta_t^0\right\|_1+\lambda\left\|\beta_t^0\right\|_1$$

We need to show that

$$(a)\ \max_{i\leq N}\left(\frac{1}{H}\left|\sum_{|j-t|\leq H} w_{tj}\epsilon_j x_{i,j}\right|\right)\leq\lambda_0 \text{ and } (b)\ \max_{i\leq N}\left(\frac{2}{H}\left|\sum_{|j-t|\leq H} w_{tj}\left(|x_{i,j}|-E\left(|x_{i,j}|\right)\right)\right|\right)\leq\lambda_0 \tag{13}$$

with probability approaching one. We leave this to the end of the proof. By our assumptions, $\max_{i\leq N}\left(\frac{2}{H}\sum_{|j-t|\leq H} w_{tj}E\left(|x_{i,j}|\right)\right)=O(1)$. Then,

$\max_{|j-t|\leq H}\left\|x_j'\left(\beta_j^0-\beta_t^0\right)\right\|\max_{i\leq N}\left(\frac{2}{H}\sum_{|j-t|\leq H} w_{tj}E\left(|x_{i,j}|\right)\right)=O_p\left(\frac{H}{T}\right)$. Then, setting $\lambda\geq 3\lambda_0$ and, by (15), $\lambda_0=(\log N)^2 H^{-\frac{1}{2}}+\frac{H}{T}$, and noting that $\left\|\hat{\beta}_t-\beta_t^0\right\|_1\leq\left\|\hat{\beta}_t\right\|_1+\left\|\beta_t^0\right\|_1$ gives $\frac{1}{H}\sum_{j=1}^T w_{tj}\left(x_j'\left(\beta_t^0-\hat{\beta}_t\right)\right)^2\leq 3\lambda\left\|\beta_t^0\right\|_1$with probability approaching 1, proving consistency, under (13).

The final matter is to show (13). We analyse (a) in (13). (b) can be analysed similarly. We note that

$$\Pr\left(\left|\sum_{j=1}^T w_{tj}\epsilon_j x_{i,j}\right|>\lambda_T\right)\leq\Pr\left(\left|\sum_{|j-t|\leq H} w_{tj}\epsilon_j x_{i,j}\right|>\lambda_T/2\right)+\Pr\left(\left|\sum_{|j-t|>H} w_{tj}\epsilon_j x_{i,j}\right|>\lambda_T/2\right) \tag{14}$$

We focus on the first term on the RHS of (14) as required by (13). The second term can be shown to be of lower order than the first term. Under our mixing assumption we have, using Theorem 3.5 of White and Wooldridge (1991), that for some $C>0$ that can be made sufficiently large by choosing appropriately the constants in (2), $\Pr\left(\frac{1}{H}\left|\sum_{j=1}^T w_{tj}\epsilon_j x_{i,j}\right|>\lambda_T/2\right)\leq\exp\left[\frac{-C\lambda_T^{\left(\frac{s}{s+2}\right)}}{H^{\frac{s(\delta-1)}{2(s+2)}}}\right]$. Further, $\Pr\left(\max_i\frac{1}{H}\left|\sum_{|j-t|\leq H} w_{tj}\epsilon_j x_{i,j}\right|>\lambda_T\right)\leq\sum_i\Pr\left(\frac{1}{H}\left|\sum_{|j-t|\leq H} w_{tj}\epsilon_j x_{i,j}\right|>\lambda_T\right)$. Note $\left(H^{\frac{s(\delta-1)}{2(s+2)}}\right)^{\left(\frac{s+2}{s}\right)}=H^{\frac{(\delta-1)}{2}}$. So

$$\lambda_T=(\log N)^{\left(\frac{s+2}{s}\right)}H^{\frac{(\delta-1)}{2}} \tag{15}$$

gives $N\exp\left[\frac{-C\lambda_T^{\left(\frac{s}{s+2}\right)}}{H^{\frac{s(\delta-1)}{2(s+2)}}}\right]=N\exp\left[\frac{-C(\log N)H^{\left(\frac{(\delta-1)}{2}\right)\left(\frac{s}{s+2}\right)}}{H^{\frac{s(\delta-1)}{2(s+2)}}}\right]$. Since s is not known an operational solution is to use $\lambda_T=\log N^2 H^{-\frac{1}{2}}$.■

		$\kappa = 2$					$\kappa = 4$				
	T	200	400	800	1600	3200	200	400	800	1600	3200
In-sample	TV	27.71	22.26	17.21	13.74	10.47	39.52	31.96	24.89	19.99	15.26
	Fix.	25.63	25.72	25.29	25.78	25.68	29.96	28.34	26.91	26.94	26.24
Out-of-sample	TV	3.036	2.836	2.466	2.250	2.350	5.617	5.496	4.765	4.451	4.585
	Fix.	3.517	3.508	3.051	2.953	3.336	5.637	5.637	4.867	4.746	5.387

Table 1: Simulation results with *iid* covariates. The row labeled "In-sample" reports the in-sample MSE for β_t averaged over all β_t's. The row labeled "Out-of-sample" reports the out-of-sample MSE for one-step ahead forecasts.

		Diebold-Mariano				Diebold-Mariano		
	MSE	LASSO	TV AR(1)	AR(1)	MAD	LASSO	TV AR(1)	AR(1)
TV LASSO	0.0630	-1.316	-1.104	-1.430	0.1701	-1.972**	-0.902	-2.133**
Const LASSO	0.0744		-0.050	-0.071	0.1874		1.256	-0.220
TV AR(1)	0.0747	0.050		-0.043	0.1787	-1.256		-1.815*
AR(1)	0.0749	0.071	0.043		0.1890	0.220	1.815*	

Table 2: Out-of-sample forecast accuracy for CPI inflation. The table reports the MSE and MAD, and the associated Diebold-Mariano test statistics for the null hypothesis of equal predictive accuracy. A negative value of the Diebold-Mariano statistic indicates that the model in the row performs better than the model in the column. *,** denote a statistically significant difference in predictive accuracy at the 10% and 5% level, respectively.